SociaRI: Enhancing Social Robot Acceptance through Ontology-Based Context Modeling and LLM-Driven Dialogue

Luca Greco[1][0000−0003−0246−2842], Francesco Rosa[1][0009−0001−2375−4839], Alessia Saggese[1✉][0000−0003−4687−7994], and Mario Vento[1][0000−0002−2948−741X]

Department of Information and Electrical Engineering and Applied Mathematics
University of Salerno, 84084 Fisciano (SA), Italy
{lgreco,frosa,asaggese,mvento}@unisa.it

Abstract. Social robots must balance fluent dialogue generation with physical and social grounding to achieve user acceptance. While Large Language Models (LLMs) provide powerful conversational capabilities, they lack environmental context and often produce hallucinations that compromise user trust. This paper proposes SociaRI, an ontology-based architecture that grounds LLM-driven dialogue with structured environmental and user context, dynamically populated by Vision-Language Models (VLMs). Our hybrid symbolic-neural approach enables physically grounded interactions with semantic coherence, personalization, and contextual awareness. We validate the system on the ARI social robot through real-world, unstructured interactions, demonstrating that combining ontological knowledge with LLM generation enhances robot behavior in authentic human-robot engagement. Our findings show that social acceptance depends not only on language fluency but also on the robot's ability to maintain context, personalize responses, and act consistently as a grounded agent.

Keywords: Social Robotics, Ontology, Knowledge Graphs, Large Language Models, Vision Language Models, Human-Robot Interaction, Context-Aware Dialogue

1 Introduction

Nowadays, robots are increasingly deployed to interact with humans in both industrial and social environments. In industrial settings, robots collaborate with human operators to perform complex tasks and provide assistance on production lines [4], while in social contexts, they engage with people at trade shows and public events to facilitate human connection and information exchange [7]. As robots transition across these diverse domains, they face a critical challenge: developing natural, context-aware, and meaningful interaction capabilities that enhance user experience and trust. This is particularly important for social robots,

[1]These authors contributed equally to this work.

which are increasingly expected to engage humans in ways that feel intuitive and responsive. Such sophisticated interaction abilities are essential to ensure social acceptance and support long-term usability in real-world deployments [22,12].

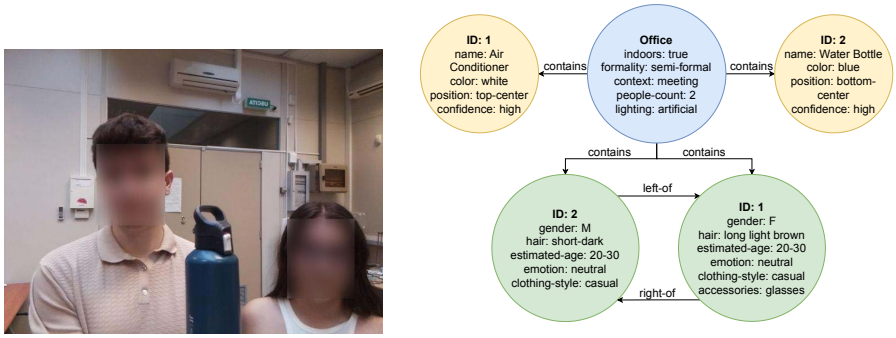
Traditional dialogue systems for social robots rely heavily on predefined scripts, which often result in repetitive or shallow interactions that diminish user engagement over time [26]. To overcome these limitations, recent research has turned to Large Language Models (LLMs). In particular, LLMs increasingly prove to function effectively as the cognitive core of robotic systems, equipping them with a vast repository of commonsense knowledge that was previously difficult for autonomous agents to acquire [30]. This cognitive layer enables robots to translate ambiguous human commands (such as "prepare breakfast") into structured, high-level task plans and subsequent low-level motion sequences. Furthermore, by leveraging modern multimodal frameworks that associate text with visual and auditory inputs, these systems allow robots to achieve a more nuanced understanding of user intent within complex physical environments. This cognitive layer helps robots understand vague or incomplete user utterances and translate them into clear dialogue intentions and responses. By using multi-modal context, such as visual and auditory cues [6], it can adjust its replies, handle conversation topics more naturally, and keep the interaction coherent in everyday situations. While traditional spoken dialogue systems were primarily designed for dyadic, one-on-one interactions, recent research has utilized LLMs to facilitate sophisticated multi-party conversations [1,5]. This evolution is critical for robots deployed in public spaces (such as museums, libraries, or hospitals) where they must address the social dynamics of interacting with groups. Within these contexts, LLMs manage complex dialogue tasks, including turn-taking arbitration, answering in-domain questions via grounded knowledge, and responding to out-of-domain requests like jokes or quizzes. Additionally, these LLM-based systems enhance accessibility by generating human-like clarification requests when a user pauses mid-sentence, a feature particularly beneficial for interacting with individuals with cognitive impairments, such as dementia.

However, LLMs alone lack grounding in the robot's physical or semantic environment, which often leads to context-induced errors or hallucinations [24], ultimately compromising user trust and acceptance in real-world scenarios. Moreover, the absence of contextual and user-specific information prevents the robot from adapting its responses to the individuals it interacts with, limiting its ability to deliver socially appropriate and personalized behaviors, an aspect known to be crucial for social acceptability. These limitations highlight the growing need to enrich LLMs with structured contextual knowledge that anchors their reasoning to the robot's surroundings and to the characteristics of the people involved in the interaction. In similar domains such as social networks, ontologies demonstrate how structured semantic models can unify heterogeneous user data and support powerful profiling and personalization mechanisms. By formally representing interests, activities, roles, and locations, ontologies make it possible to merge posts, reactions, and cross-platform behaviors into coherent user profiles [17]. Domain and linguistic ontologies classify text and shared content

into meaningful categories [21], while ontology-driven digital profiling organizes diverse artifacts such as images, videos, and social traces into interpretable user characteristics [10]. Social knowledge graphs further enrich these representations by linking users, content, entities, and interactions, enabling advanced reasoning for recommendation, influence detection, and community analysis [27]. These capabilities illustrate how semantic models can capture user interests, context, and relationships; such insights can be directly transferable to social robotics, where similar ontology-based profiling could inform adaptive behavior and socially aware decision-making [15]. In fact, ontological representations can help robots to track entities, their relations, and the user's evolving state, enabling more consistent, interpretable, and user-tailored dialogue management [25]. Moreover, they can also be effective for enriching personalized interactions [8], supporting task planning [16], and mitigating ambiguity in user instructions [18]. The integration of symbolic knowledge with LLM-driven generation thus emerges as a promising path toward achieving both flexibility and reliability.

At the perceptual level, Vision-Language Models (VLMs) provide a powerful tool to extract semantically rich contextual information from the environment. They can support user modeling, scene understanding, and behavioral adaptation [11]. Recent studies demonstrate that VLMs can identify salient visual cues, detect social entities, and assist in context-sensitive decision-making within HRI scenarios [29,13]. This aligns with recent directions emphasising the need for user-aware conversational systems [23], diversity-sensitive interaction [9], and proactive dialogue management [19]. Using these multimodal capabilities to populate an ontology could provide a structured and verifiable substrate for grounding robot perception into knowledge-driven interaction policies. Starting from these assumptions, in this work, we propose an ontology-based knowledge graph tailored for social robotics, designed to support structured, user-specific context modelling and dialogue control in LLM-powered robots. The ontology is dynamically populated using multimodal perception pipelines, including VLMs capable of extracting attributes such as user appearance, affective cues, and scene context. By combining symbolic knowledge populated by VLM, with the generative strengths of LLMs, we aim to achieve interactions that are more adaptive and socially grounded.

As a case study, we demonstrate our approach on the ARI robot by PAL Robotics. We call the proposed solution SociaRI, the Social ARI Robotic Platform. Inspired by similar solutions integrating LLMs into embodied agents such as Pepper, Furhat, and Mini [22,20,28], our system highlights how the ARI robot can benefit from a unified ontology-LLM-VLM pipeline to deliver richer conversational experiences. This integration supports turn-level natural language generation, paraphrasing, empathy, small talk, topic expansion, and context-sensitive utterance selection, while maintaining semantic coherence through knowledge-based grounding. Overall, this contribution advances the growing body of work on hybrid symbolic- neural architectures for HRI by presenting a structured approach to enhance LLM-powered social robots with ontological knowledge and multimodal perception. Our findings reinforce that social acceptance depends not



(a) (b)

Fig. 1. Illustration of the proposed Knowledge Graph (KG). (a) Experimental setup showing two participants interacting with the robot. (b) Corresponding KG instantiation, where the **Environment** node (*Office*) encodes scene-level attributes (formality, lighting, people count), **Person** nodes store VLM-extracted individual attributes (gender, hair, estimated age, emotion, clothing style), and **Object** nodes capture detected items with their visual properties (e.g., the *Air Conditioner* and *Water Bottle* with color, position, and confidence). Spatial relationships (e.g., *left-of*, *right-of*, *contains*) are encoded as directed edges, capturing the geometric configuration of entities within the scene and enabling context-aware Human-Robot Interaction.

only on fluent language generation but also on the robot's ability to maintain context, personalize interaction, and act consistently within shared environments [22,14].

The remainder of this paper is organized as follows. Section 2 introduces the proposed Knowledge Graph, along with its VLM-based generation. Section 3 details the proposed SociaRI architecture. Section 4 presents and discusses the results. Finally, Section 5 concludes the paper and outlines directions for future research.

2 Knowledge Graph

In this paragraph we are going to describe the proposed *Knowledge Graph* (KG) and the way how we built it. Specifically, an ontology can be represented as a graph; a graph $\mathcal{G}$ can be represented as a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges. An example of the proposed KG instantiation is shown in Fig. 1, where the VLM processes the scene, extracting person-level attributes, object-level properties, and scene-level descriptors that are used to populate the graph. In our application, vertices can represent 3 different kind of entities: (•) **Environment**, this node is characterized by a name, which represent the kind of environment (e.g., office, room, lab, etc. . .), and a set of attributes related to a textual description of the environment generated by the VLM. As visible in Fig. 1(b), the *Office* node encodes attributes such

Prompts to populate the Knowledge Graph

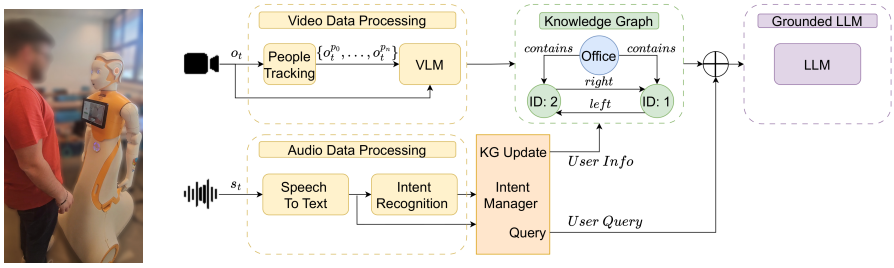
Environment Prompt: : Analyze the full scene. Return JSON with: `location`, `indoors`, `context`, `people_count`, `lighting`, `confidence`.
Person-Light Prompt: Analyze visible person. Return JSON with: `apparent_gender`, `estimated_age_range`, `emotion`, `confidence`.
Person-Deep Prompt: Analyze person in detail. Return JSON with: `hair`, `glasses`, `clothing_style`, `dominant_clothing_colors`, `accessories`, `confidence`.
Object Prompt: List all non-person objects. Return JSON array with per-object: `name`, `category`, `color`, `size`, `position`, `confidence`.
Relationship Prompt: Analyze two highlighted people (GREEN and RED boxes). Return JSON with: `subject_id`, `reference_id`, `space_dist` (left/right/overlapping), `confidence`. Base analysis on visible body language and proximity.

Fig. 2. List of prompts used to populate the Knowledge Graph.

as formality, lighting conditions, and the number of people present in the scene; (•) **Person**, this node represents a person, it is characterized by the track-id and it is enhanced by a set of attributes related to the people (e.g., age, emotion, gender and characteristics) extracted by the VLM during the inference. For instance, in Fig. 1(b), two *Person* nodes are instantiated, each storing individually inferred attributes such as estimated age, hair, emotion, and clothing style; (•) **Object**, this node represents an object detected by the VLM. It is characterized by an object-id and enhanced through visual characteristics extracted by the VLM (e.g., color, position, confidence). As shown in Fig. 1(b), two *Object* nodes are present in the scene: the *Air Conditioner* (color: white, position: top-center) and the *Water Bottle* (color: blue, position: bottom-center), both connected to the *Office* node via *contains* edges. About edges, they encode relationships between entities and allow us to capture the structural dependencies within the scene. We define two primary categories of relationships: (•) **Spatial relationships** encode the physical proximity and geometric configuration between entities. These relationships capture their relative positions and orientations in the environment, such as *left, right, in front of, above, contains*, and similar spatial predicates. As shown in Fig. 1(b), spatial edges such as *left-of*, *right-of*, and *contains* connect the entity nodes, reflecting the geometric layout of the scene, linking both *Person* and *Object* nodes to the *Environment*; (•) **Social relationships** represent how entities are socially or functionally connected to each other. These include connections based on roles, interactions, and associations, such as *friends, colleagues, supervisor*, or task-based relationships. Together, these two relationship types form a comprehensive framework for representing both the physical environment and the social context necessary for effective Human-Robot Interaction.

2.1 Graph Generation

In this section, we describe the procedure used to populate the Knowledge Graph. The graph is automatically generated by prompting a VLM, specifically Qwen2.5-VL-3B [2]. Since the ontology comprises three entity types, Environment, Person, and Object, and two relationship types, Spatial and Social, we



(a) (b)

Fig. 3. (a) Experimental setup with a participant interacting in front of the ARI robot. (b) Proposed high-level control architecture for the Physically-Grounded LLM for Human-Robot Interaction. The *Video Data Processing* pipeline processes each camera frame o_t through a *People Tracking* module, which detects and tracks individuals producing per-person crops, and a *VLM* module, which generates and updates the KG from both the scene and the detected people. In parallel, the *Audio Data Processing* pipeline transcribes the audio signal via *Speech-to-Text* and classifies it through *Intent Recognition*; if a grounded query is detected, it is concatenated with the KG and passed to the *Grounded LLM* to produce contextually informed responses.

design five distinct prompts to automatically extract relevant information from the input image, as summarized in Figure 2.

Moreover, leveraging natural language communication, we integrate a complementary mechanism that allows users to dynamically update the Knowledge Graph during interaction with the robot. Specifically, users can modify their own attributes (e.g., preferred city, favorite food, dietary restrictions, allergies) and update relationships between themselves and other entities in the scene. In this way, the robot maintains an accurate and up-to-date representation of the social and personal context, ensuring that its interactions remain personalized and contextually appropriate. This user-driven update mechanism enhances system transparency, increases user engagement, and enables the robot to continuously adapt to evolving human preferences and social dynamics without requiring manual intervention.

3 SociARI architecture

We validate the proposed architecture on the social robot ARI, developed by PAL Robotics (Figure 3a). The robot is equipped with a general-purpose 8MP RGB camera mounted on its head, which is used to capture images of the environment, and a ReSpeaker microphone array for vocal interaction. Figure 3b illustrates the high-level control architecture, highlighting the most relevant components and the main information flows pertaining to the proposed system.

The current frame o_t delivered by the robot's camera is processed by two modules: (•) The **People Tracking** module, which takes o_t as input, detects

people in the image, and tracks them over time, returning for each person p_i a cropped image corresponding to the detected bounding box, denoted $o_t^{p_i}$; (•) The **VLM** module, which takes as input both the overall scene observation o_t and the set of detected people $\{o_t^{p_0}, \ldots, o_t^{p_n}\}$, and generates the KG as described in Section 2.

Since the goal of the proposed architecture is to enable physically grounded interaction, we also implement an Audio Data Processing Pipeline. Given the current audio signal s_t as input, it first transcribes it into text via a **Speech-to-Text** module, which is then passed to an **Intent Recognition** module to determine whether the text contains a *grounded query*. If the user asks something about the environment or the tracked people, the query is concatenated with the Knowledge Graph and fed to the chosen LLM, which can then answer the question informed by the environmental context provided by the KG.

4 Experiments

In this section, we describe the experimental setup. Specifically, SociaRI is compared against the ROMAN baseline [7], described in Section 4.1, with the goal of addressing the following research question: *How does a social robot behave during real-world, unstructured interactions when augmented with contextual information from the Knowledge Graph?*. The evaluation metrics are described in Section 4.2, and the results are presented in Section 4.3.

4.1 Baseline

As previously mentioned, SociaRI is compared against ROMAN [7], a state-of-the-art social robot architecture (Figure 4) composed of three main components: (•) **Orchestrator**, which routes outputs based on detected user intent. For queries about people attributes, it selects corresponding estimation modules; otherwise, a Fallback mechanism invokes the VLM directly on the current scene observation o_t; (•) **Video Data Processing**, which performs face and object detection on o_t, but produces isolated per-module outputs rather than a structured semantic representation; (•) **Audio Data Processing**, which comprises Speech-to-Text and Intent Recognition modules. Unlike the proposed system, recognized intent directly commands the Orchestrator rather than conditioning a Knowledge Graph query.

Despite providing a functional foundation for social robot interaction, this architecture exhibits several limitations that the proposed system aims to address: (•) **Error Propagation**, since ROMAN follows a modular pipeline design, an error in any individual module can propagate to downstream components, ultimately degrading the overall interaction quality. Although the ontology-based Knowledge Graph generated by the VLM is also subject to perceptual errors, our approach mitigates this issue by allowing the robot to dynamically update and correct node information through voice interaction, enabling a form of in-session error recovery; (•) **Lack of Environmental Grounding**, ROMAN lacks a

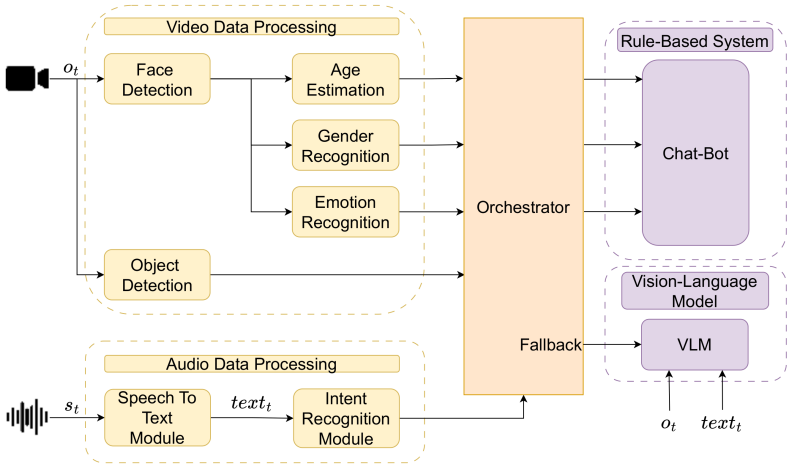


Fig. 4. Architecture of the ROMAN baseline [7], where the orchestrator processes soft-biometric, objects, and user intent to select between a rule-based chatbot and a VLM module.

structured and unified representation of the environment, limiting its ability to support grounded interactions. Specifically, it cannot reason about contextual aspects such as the overall scene, spatial relationships among individuals, or the social dynamics emerging from their interactions. In contrast, the proposed ontology-based Knowledge Graph explicitly encodes this information, enabling the LLM to generate responses that are semantically grounded in the robot's perceived physical and social context.

4.2 Evaluation Metrics

We assessed system performance using a 15-item questionnaire administered on a 5-point Likert scale. Items Q4–Q13 were adapted from the *Godspeed Questionnaire* [3], while Q14–Q18 were custom-designed to evaluate features specific to the proposed system. The questionnaire measures four constructs:

- **Perceived Intelligence (Q4–Q8)**: captures the user's perception of the robot's cognitive abilities through five bipolar adjective pairs: *Incompetent–Competent*, *Ignorant–Knowledgeable*, *Irresponsible–Responsible*, *Unintelligent–Intelligent*, and *Foolish–Sensible*.
- **Likability (Q9–Q13)**: assesses the robot's affective impact and perceived social presence through five bipolar adjective pairs: *Dislike–Like*, *Unfriendly–Friendly*, *Unkind–Kind*, *Unpleasant–Pleasant*, and *Awful–Nice*.
- **Memory (Q14–Q15)**: evaluates the robot's ability to maintain contextual coherence throughout the interaction. Q14 asks *"Did the conversation maintain a logical thread from start to finish?"*, rated from *Incoherent* to *Coherent*.

Q15 asks *"Did the final answers take into account the context built up in the earlier turns?"*, rated from *Disconnected* to *Contextualized*.
- **Personalization (Q16–Q18)**: measures the effectiveness of the Knowledge Graph in tailoring responses to the user. Q16 asks *"Was the conversation personalized on the basis of your biometric information (e.g., age, gender, ethnicity, emotion)?"*. Q17 asks *"Was the conversation personalized on the basis of the personal information you shared with the robot during the dialogue?"*. Q18 asks, more generally, *"Overall, did the conversation feel tailored to you?"*

Questionnaire responses were analyzed using the following pipeline: (i) **Sample Size Justification**, an *a priori* power analysis using `G*Power` indicated a minimum of 22 paired observations to detect a large effect size ($d = 0.8$) at a significance level of $\alpha = 0.05$ with power $1 - \beta = 0.95$. Our sample of 46 participants exceeds this requirement; (ii) **Measurement**, for each construct, composite scores were computed as the mean of the corresponding items. Internal consistency was assessed using Cronbach's α, with values $\alpha \geq 0.70$ considered acceptable; (iii) **Hypothesis Testing**, given the ordinal nature of Likert-scale data and the within-subjects design, baseline and Knowledge Graph-augmented conditions were compared using a one-tailed paired Wilcoxon signed-rank test ($\alpha = 0.05$); (iv) **Effect Sizes and Power,** practical significance was quantified using paired Cohen's d_z (computed from the standard deviation of paired differences), with conventional thresholds: small ($d_z \approx 0.2$), medium ($d_z \approx 0.5$), and large ($d_z \geq 0.8$). Post-hoc statistical power ($1 - \beta$) was computed for each test, with values ≥ 0.80 considered adequate.

4.3 Results

In this section, we present the results obtained from the evaluation protocol described above. Overall, we conducted a within-subjects user study with 46 participants, following established protocols from prior work [7,6]. Each participant interacted with both the baseline and proposed systems in separate 10-minute sessions under naturalistic conditions. Participants were encouraged to share personal information and ask questions about the environment. At the end of each session, they completed the questionnaire described above.

Overall Findings Table 1 reports the results of our within-subjects A/B study comparing the ROMAN baseline architecture with the proposed SociaRI architecture. The results reveal statistically significant improvements across all four constructs, with the largest gains observed in Memory and Personalization. Effect sizes ranged from medium to large ($d_z = 0.60$–0.87), with Memory and Personalization exhibiting the strongest effects ($d_z \geq 0.86$). Statistical power was no lower than 0.96 for any construct, exceeding 0.99 for three of the four. Internal consistency was also high across both conditions for all constructs (Cronbach's $\alpha \geq 0.86$; Table 2), supporting the reliability of the questionnaire. Overall, these results provide strong evidence for the superiority of the proposed SociaRI

Table 1. Within-subjects A/B study results. For both the ROMAN and the proposed SociaRI architectures we report means (M) and standard deviations (SD) across 4 constructs: Perceived Intelligence (P.I.), Likability (L.), Memory (M.), Personalization (P.). Wilcoxon signed-rank tests were used to assess statistical significance, reporting the test statistic (W), p-value, standardized effect size (r), Cohen's d_z, and statistical power ($1-\beta$)

	ROMAN		**SociaRI (Prop.)**						
	M	SD	M	SD	W	p	r	d_z	$1-\beta$
P.I.	3.53	0.84	**4.25**	**0.60**	109.5	3.1×10^{-5}	0.69	0.73	>0.99
L.	3.57	0.88	**4.16**	**0.69**	179.5	1.1×10^{-3}	0.58	0.60	0.96
M.	2.98	0.97	**4.23**	**0.88**	79.0	7.0×10^{-6}	0.74	0.87	>0.99
P.	3.10	0.94	**4.16**	**0.85**	90.0	4.0×10^{-6}	0.73	0.86	>0.99

architecture, with Memory and Personalization emerging as the primary contributors to the observed improvements.

Construct-Specific Findings Building on the aggregate results reported above, we examine each of the four constructs individually to characterize the specific dimensions along which the SociaRI architecture outperformed the ROMAN baseline.

Perceived Intelligence. Participants perceived the SociaRI architecture as significantly more intelligent than ROMAN ($M = 4.25$, $SD = 0.60$ vs. $M = 3.53$, $SD = 0.84$; $W = 109.5$, $p = 3.1 \times 10^{-5}$, $r = 0.69$, $d_z = 0.73$, $1 - \beta > 0.99$), corresponding to a medium-to-large effect. This suggests that grounding LLM responses in a structured semantic representation enhanced users' perception of the robot's cognitive capabilities, even though the underlying LLM remained constant across conditions. The robot's ability to reference environmental context and maintain logical consistency likely contributed to this perception.

Likability. The SociaRI architecture was also rated as significantly more likeable and socially present than ROMAN ($M = 4.16$, $SD = 0.69$ vs. $M = 3.57$, $SD = 0.88$; $W = 179.5$, $p = 1.1 \times 10^{-3}$, $r = 0.58$, $d_z = 0.60$, $1 - \beta = 0.96$). Although this was the smallest effect size among the four constructs, the improvement remained highly significant, with power confirming adequate sensitivity to detect the observed effect. This finding suggests that contextually grounded dialogue contributes meaningfully to users' affective engagement with the robot, fostering a more positive and personable interaction experience.

Memory. The most pronounced improvement was observed in the Memory construct, which measures the robot's ability to maintain logical coherence across the interaction. SociaRI achieved a markedly higher mean than ROMAN ($M = 4.23$, $SD = 0.88$ vs. $M = 2.98$, $SD = 0.97$; $W = 79.0$, $p = 7.0 \times 10^{-6}$, $r = 0.74$, $d_z = 0.87$, $1 - \beta > 0.99$), representing the largest effect size observed across all constructs. This result directly supports our core hypothesis: by dynamically populating an ontology-based Knowledge Graph, the SociaRI

Table 2. Cronbach's α internal consistency coefficients for the four measurement constructs in both the ROMAN and in the proposed SociaRI conditions.

	Perceived Intelligence	**Likability**	**Memory**	**Personalization**
α_{ROMAN}	0.92	0.92	0.89	0.87
α_{SociaRI}	0.86	0.88	0.93	0.93

architecture enables the LLM to generate responses explicitly grounded in prior interaction context and environmental state, thereby reducing hallucinations and enhancing narrative coherence.

Personalization. The Personalization construct showed similarly substantial improvement, with SociaRI ($M = 4.16$, $SD = 0.85$) significantly outperforming ROMAN ($M = 3.10$, $SD = 0.94$; $W = 90.0$, $p = 4.0 \times 10^{-6}$, $r = 0.73$, $d_z = 0.86$, $1 - \beta > 0.99$). Participants reported that the Knowledge Graph-augmented system better tailored responses based on their biometric information, personal attributes shared during interaction, and individual preferences. This demonstrates that the user-driven update mechanism and semantic representation of personal attributes significantly enhance the robot's capacity for adaptive, user-centered interaction.

Internal Consistency Table 2 reports Cronbach's α coefficients for both conditions. All constructs in both conditions achieved $\alpha \geq 0.86$, substantially exceeding the conventional threshold for acceptable internal consistency ($\alpha \geq 0.70$) and demonstrating good to excellent reliability of the measurement scales across both ROMAN and SociaRI conditions.

Interpretation and Implications Augmenting LLM-based social robots with a dynamically populated Knowledge Graph significantly enhances user-perceived quality across all dimensions, with particularly large improvements in Memory and Personalization ($d_z = 0.87$ and $d_z = 0.86$). These gains directly address a fundamental LLM limitation, the lack of persistent, structured context for maintaining coherent agent identity, by maintaining a unified semantic representation of environmental entities and user attributes that enables contextually appropriate, non-hallucinatory responses grounded in factual knowledge rather than probabilistic generation alone. The effect sizes are practically meaningful, supporting the premise that social acceptance in human-robot interaction depends critically on the robot's ability to maintain context and act as a coherent agent. The proposed SociaRI architecture achieves this through a hybrid symbolic-neural pipeline that anchors LLM generation in structured semantic knowledge, effectively bridging the gap between neural fluency and symbolic groundedness.

5 Conclusions

In this work, we presented SociaRI, a socially aware robotic platform integrating an ontology-based Knowledge Graph with multimodal perception and

LLM-driven dialogue. Our hybrid symbolic–neural architecture, in which Vision-Language Models dynamically populate a structured ontology, addresses a central challenge in social robotics: enabling contextually grounded, personalized, and semantically coherent interactions rather than merely linguistically fluent exchanges. Through real-world interaction scenarios, we demonstrated that grounding LLM responses in a dynamically updated Knowledge Graph significantly improves context maintenance, hallucination avoidance, and user-specific personalization. SociaRI exhibited substantially improved performance over the ROMAN baseline across all four measurement constructs, with particularly large gains in Memory and Personalization ($d_z = 0.87$ and $d_z = 0.86$), confirming the practical value of hybrid symbolic–neural architectures for social robotics.

References

1. Addlesee, A., Cherakara, N., Nelson, N., Hernández-García, D., Gunson, N., Sieińska, W., Romeo, M., Dondrup, C., Lemon, O.: A multi-party conversational social robot using llms. In: ACM/IEEE International Conference on Human-Robot Interaction. pp. 1273 – 1275 (2024)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
3. Bartneck, C.: Godspeed questionnaire series: Translations and usage. In: International handbook of behavioral health assessment, pp. 1–35. Springer (2023)
4. De Simone, G., Greco, A., Rosa, F., Saggese, A., Vento, M.: Context-aware data augmentation for enhanced speech command recognition in industrial environments. Scientific Reports **15**(1), 17445 (2025)
5. De Simone, G., Greco, L., Saggese, A., Vento, M.: Robot to human interaction with multi-modal conversational engagement. In: 2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR). pp. 1–8 (2025)
6. De Simone, G., Greco, L., Saggese, A., Vento, M.: Integrating visual and audio cues for emotion and gender recognition: A multi modal and multi task approach. Information Fusion **130** (2026)
7. De Simone, G., Saggese, A., Vento, M.: Empowering human interaction: A socially assistive robot for support in trade shows. In: 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). pp. 901–908. IEEE (2024)
8. Grassi, L., Recchiuto, C.T., Sgorbissa, A.: Knowledge triggering, extraction and storage via human–robot verbal interaction. Robotics and Autonomous Systems **148** (2022)
9. Grassi, L., Recchiuto, C.T., Sgorbissa, A.: Enhancing llm-based human-robot interaction with nuances for diversity awareness. In: IEEE International Workshop on Robot and Human Communication, RO-MAN. pp. 2287 – 2294 (2024)
10. Grigaliunas, S., Bruzgiene, R., Venckauskas, A.: Ontology-driven digital profiling for identification and linking evidence across social media platform. IEEE Access **11**, 111672 – 111691 (2023)

11. Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., Xu, R., Xu, S.: Multimodal fusion and vision–language models: A survey for robot vision. Information Fusion **126** (2026)
12. Irfan, B., Kuoppamäki, S.M., Hosseini, A., Skantze, G.: Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. Autonomous Robots **49**(1) (2025)
13. Jiang, C., Dehghan, M., Jägersand, M.: Understanding contexts inside robot and human manipulation tasks through vision-language model and ontology system in video streams. In: IEEE International Conference on Intelligent Robots and Systems. pp. 8366 – 8372 (2020)
14. Mahieu, C., Ongenae, F., de Backere, F., Bonte, P., de Turck, F.D., Simoens, P.: Semantics-based platform for context-aware and personalized robot interaction in the internet of robotic things. Journal of Systems and Software **149**, 138 – 157 (2019)
15. Manzoor, S., Rocha, Y.G., Joo, S., Bae, S., Kim, E.j., Joo, K., Kuc, T.: Ontology-based knowledge representation in robotic systems: A survey oriented toward applications. Applied Sciences (Switzerland) **11**(10) (2021)
16. Molina, V., Ruiz-Celada, O., Suarez, R., Rosell, J., Zaplana, I.: Robot situation and task awareness using large language models and ontologies. In: 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). pp. 96–103. IEEE (2025)
17. Moshkin, V.: The approach to building a graph knowledge base using social media data. In: 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT). pp. 1–6. IEEE (2020)
18. Nakajima, H., Miura, J.: Combining ontological knowledge and large language model for user-friendly service robots. In: IEEE International Conference on Intelligent Robots and Systems. pp. 4755 – 4762 (2024)
19. Niu, Z., Xie, Z., Cao, S., Lu, C., Ye, Z., Xu, T., Liu, Z., Gao, Y., Chen, J., Xu, Z., Wu, Y., Hu, Y.: Part: Enhancing proactive social chatbots with personalized real-time retrieval. In: SIGIR 2025 - Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 4269 – 4274 (2025)
20. Onorati, T., Castro-González, Á., Del Valle, J.C., Díaz, P., Castillo, J.C.: Creating personalized verbal human-robot interactions using llm with the robot mini. In: International conference on ubiquitous computing and ambient intelligence. pp. 148–159. Springer (2023)
21. Peña, P., Del Hoyo, R., Vea-Murguía, J., González, C., Mayo, S.: Automatic ontology user profiling for social networks from urls shared. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8109 LNAI**, 168 – 177 (2013)
22. Pinto-Bernal, M.J., Biondina, M., Belpaeme, T.: Designing social robots with llms for engaging human interaction. Applied Sciences (Switzerland) **15**(11) (2025)
23. Rahimi, H., Cattoni, J., Beghili, M., Abrini, M., Khoramshahi, M., Pino, M., Chetouani, M.: Reasoning llms for user-aware multimodal conversational agents. In: IEEE International Workshop on Robot and Human Communication, RO-MAN. pp. 443 – 448 (2025)
24. Ranasinghe, N., Mohammed, W.M., Stefanidis, K., Martínez Lastra, J.L.L.: Large language models in human-robot collaboration with cognitive validation against context-induced hallucinations. IEEE Access **13**, 77418 – 77430 (2025)

25. Saracco, A., Lillo, A., Stranisci, M.A., Gena, C.: Human robot interaction through an ontology-based dialogue engine. In: ACM/IEEE International Conference on Human-Robot Interaction. pp. 940 – 944 (2024)
26. Sevilla-Salcedo, J., Fernádez-Rodicio, E., Martín-Galván, L., Castro-González, Á., Castillo, J.C., Salichs, M.A.: Using large language models to shape social robots' speech. International Journal of Interactive Multimedia and Artificial Intelligence (2023)
27. Shen, Y., Jiang, X., Li, Z., Wang, Y., Xu, C., Shen, H., Cheng, X.: Uniskgrep: A unified representation learning framework of social network and knowledge graph. Neural Networks **158**, 142 – 153 (2023)
28. Sievers, T., Russwinkel, N.: Retrieving memory content from a cognitive architecture by impressions from language models for use in a social robot. Applied Sciences (Switzerland) **15**(10) (2025)
29. Song, D., Liang, J., Payandeh, A., Raj, A.H., Xiao, X., Manocha, D.: Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. IEEE Robotics and Automation Letters (2024)
30. Zhang, C., Chen, J., Li, J., Peng, Y., Mao, Z.: Large language models for human–robot interaction: A review. Biomimetic Intelligence and Robotics **3**(4) (2023)